林轩田《机器学习技法》课程笔记2 -- Dual Support Vector Machine

作者：红色石头　　公众号：AI有道（id：redstonewill）

上节课我们主要介绍了线性支持向量机（Linear Support Vector Machine）。Linear SVM的目标是找出最“胖”的分割线进行正负类的分离，方法是使用二次规划来求出分类线。本节课将从另一个方面入手，研究对偶支持向量机（Dual Support Vector Machine），尝试从新的角度计算得出分类线，推广SVM的应用范围。

Motivation of Dual SVM

首先，我们回顾一下，对于非线性SVM，我们通常可以使用非线性变换将变量从x域转换到z域中。然后，在z域中，根据上一节课的内容，使用线性SVM解决问题即可。上一节课我们说过，使用SVM得到large-margin，减少了有效的VC Dimension，限制了模型复杂度；另一方面，使用特征转换，目的是让模型更复杂，减小E_{in}。所以说，非线性SVM是把这两者目的结合起来，平衡这两者的关系。那么，特征转换下，求解QP问题在z域中的维度设为$\hat{d}+1$，如果模型越复杂，则$\hat{d}+1$越大，相应求解这个QP问题也变得很困难。当$\hat{d}$无限大的时候，问题将会变得难以求解，那么有没有什么办法可以解决这个问题呢？一种方法就是使SVM的求解过程不依赖$\hat{d}$，这就是我们本节课所要讨论的主要内容。

$$\min_{b,\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$
$$\text{s. t.} \quad y_n(\mathbf{w}^T \underbrace{\mathbf{z}_n}_{\Phi(\mathbf{x}_n)} + b) \geq 1,$$
$$\text{for } n = 1, 2, \ldots, N$$

Non-Linear Hard-Margin SVM

1. $\mathrm{Q} = \begin{bmatrix} 0 & \mathbf{0}_{\tilde{d}}^T \\ \mathbf{0}_{\tilde{d}} & \mathrm{I}_{\tilde{d}} \end{bmatrix}; \mathbf{p} = \mathbf{0}_{\tilde{d}+1};$ $\mathbf{a}_n^T = y_n \begin{bmatrix} 1 & \mathbf{z}_n^T \end{bmatrix}; c_n = 1$
2. $\begin{bmatrix} b \\ \mathbf{w} \end{bmatrix} \leftarrow \mathrm{QP}(\mathrm{Q}, \mathbf{p}, \mathrm{A}, \mathbf{c})$
3. return $b \in \mathbb{R}$ & $\mathbf{w} \in \mathbb{R}^{\tilde{d}}$ with $g_{\text{SVM}}(\mathbf{x}) = \text{sign}(\mathbf{w}^T\Phi(\mathbf{x}) + b)$

比较一下，我们上一节课所讲的Original SVM二次规划问题的变量个数是$\hat{d}+1$，有N

个限制条件；而本节课，我们把问题转化为对偶问题（'Equivalent' SVM），同样是二次规划，只不过变量个数变成N个，有N+1个限制条件。这种对偶SVM的好处就是问题只跟N有关，与$\hat{d}$无关，这样就不存在上文提到的当$\hat{d}$无限大时难以求解的情况。

Original SVM

(convex) QP of

- $\tilde{d}+1$ variables
- N constraints

'Equivalent' SVM

(convex) QP of

- N variables
- $N+1$ constraints

如何把问题转化为对偶问题（'Equivalent' SVM），其中的数学推导非常复杂，本文不做详细数学论证，但是会从概念和原理上进行简单的推导。

还记得我们在《机器学习基石》课程中介绍的Regularization中，在最小化E_{in}的过程中，也添加了限制条件：$w^Tw\leq C$。我们的求解方法是引入拉格朗日因子λ，将有条件的最小化问题转换为无条件的最小化问题：

$min\ E_{aug}(w)=E_{in}(w)+\frac{\lambda}{N}w^Tw$，最终得到的w的最优化解为：

$$\nabla E_{in}(w)+\frac{2\lambda}{N}w=0$$

所以，在regularization问题中，λ是已知常量，求解过程变得容易。那么，对于dual SVM问题，同样可以引入λ，将条件问题转换为非条件问题，只不过λ是未知参数，且个数是N，需要对其进行求解。

Regularization by Constrained-Minimizing E_{in}

$$\min_{\mathbf{w}} E_{in}(\mathbf{w}) \text{ s.t. } \mathbf{w}^T\mathbf{w}\leq C$$



Regularization by Minimizing E_{aug}

$$\min_{\mathbf{w}} E_{aug}(\mathbf{w})=E_{in}(\mathbf{w})+\frac{\lambda}{N}\mathbf{w}^T\mathbf{w}$$

- C equivalent to some $\lambda\geq 0$ by checking **optimality condition**

$$\nabla E_{in}(\mathbf{w})+\frac{2\lambda}{N}\mathbf{w}=\mathbf{0}$$

- regularization: view λ as **given parameter instead of C**, and solve 'easily'
- dual SVM: view λ's as unknown given the constraints, and **solve them as variables instead**

如何将条件问题转换为非条件问题？上一节课我们介绍的SVM中，目标是：$min\ \frac{1}{2}w^Tw$，条件是：$y_n(w^Tz_n+b)\geq 1,\ for\ n=1,2,\cdots,N$。首先，我们令拉格朗日因子为$\alpha_n$（区别于regularization），构造一个函数：

$$L(b,w,\alpha)=\frac{1}{2}w^Tw+\sum_{n=1}^{N}\alpha_n(1-y_n(w^Tz_n+b))$$

这个函数右边第一项是SVM的目标，第二项是SVM的条件和拉格朗日因子α_n的乘积。我们把这个函数称为拉格朗日函数，其中包含三个参数：b，w，α_n。

$\min_{b,\mathbf{w}}$ $\frac{1}{2}\mathbf{w}^T\mathbf{w}$

s.t. $y_n(\mathbf{w}^T\mathbf{z}_n+b)\geq 1$,

for $n=1,2,\ldots,N$

Lagrange Function

with Lagrange multipliers ~~λ_n~~ α_n,

$$\mathcal{L}(b,\mathbf{w},\alpha)=\underbrace{\frac{1}{2}\mathbf{w}^T\mathbf{w}}_{\text{objective}}+\sum_{n=1}^{N}\alpha_n\underbrace{(1-y_n(\mathbf{w}^T\mathbf{z}_n+b))}_{\text{constraint}}$$

下面，我们利用拉格朗日函数，把SVM构造成一个非条件问题：

Claim

$$\text{SVM}\equiv\min_{b,\mathbf{w}}\left(\max_{\text{all }\alpha_n\geq 0}\mathcal{L}(b,\mathbf{w},\alpha)\right)=\min_{b,\mathbf{w}}\left(\infty\text{ if violate };\ \frac{1}{2}\mathbf{w}^T\mathbf{w}\text{ if feasible}\right)$$

- any 'violating' $(b,\mathbf{w})$: $\max_{\text{all }\alpha_n\geq 0}\left(\square+\sum_n\alpha_n(\text{some positive})\right)\rightarrow\infty$
- any 'feasible' $(b,\mathbf{w})$: $\max_{\text{all }\alpha_n\geq 0}\left(\square+\sum_n\alpha_n(\text{all non-positive})\right)=\square$

该最小化问题中包含了最大化问题，怎么解释呢？首先我们规定拉格朗日因子$\alpha_n\geq 0$，根据SVM的限定条件可得：$(1-y_n(w^Tz_n+b))\leq 0$，如果没有达到最优解，即有不满足$(1-y_n(w^Tz_n+b))\leq 0$的情况，因为$\alpha_n\geq 0$，那么必然有$\sum_n\alpha_n(1-y_n(w^Tz_n+b))\geq 0$。对于这种大于零的情况，其最大值是无解的。如果对于所有的点，均满足$(1-y_n(w^Tz_n+b))\leq 0$，那么必然有$\sum_n\alpha_n(1-y_n(w^Tz_n+b))\leq 0$，则当$\sum_n\alpha_n(1-y_n(w^Tz_n+b))=0$时，其有最大值，最大值就是我们SVM的目标：$\frac{1}{2}w^Tw$。因此，这种转化为非条件的SVM构造函数的形式是可行的。

Lagrange Dual SVM

现在，我们已经将SVM问题转化为与拉格朗日因子α_n有关的最大最小值形式。已知$\alpha_n \geq 0$，那么对于任何固定的α'，且$\alpha'_n \geq 0$，一定有如下不等式成立：

for any fixed α' with all $\alpha'_n \geq 0$,

$$\min_{b,\mathbf{w}} \left(\max_{\text{all } \alpha_n \geq 0} \mathcal{L}(b, \mathbf{w}, \alpha) \right) \geq \min_{b,\mathbf{w}} \mathcal{L}(b, \mathbf{w}, \alpha')$$

对上述不等式右边取最大值，不等式同样成立：

for best $\alpha' \geq \mathbf{0}$ on RHS,

$$\min_{b,\mathbf{w}} \left(\max_{\text{all } \alpha_n \geq 0} \mathcal{L}(b, \mathbf{w}, \alpha) \right) \geq \underbrace{\max_{\text{all } \alpha'_n \geq 0} \min_{b,\mathbf{w}} \mathcal{L}(b, \mathbf{w}, \alpha')}_{\text{Lagrange dual problem}}$$

上述不等式表明，我们对SVM的min和max做了对调，满足这样的关系，这叫做Lagrange dual problem。不等式右边是SVM问题的下界，我们接下来的目的就是求出这个下界。

已知≥是一种弱对偶关系，在二次规划QP问题中，如果满足以下三个条件：

- **函数是凸的（convex primal）**
- **函数有解（feasible primal）**
- **条件是线性的（linear constraints）**

那么，上述不等式关系就变成强对偶关系，≥变成=，即一定存在满足条件的解(b, w, α)，使等式左边和右边都成立，SVM的解就转化为右边的形式。

经过推导，SVM对偶问题的解已经转化为无条件形式：

$$\max_{\text{all } \alpha_n \geq 0} \left(\underbrace{\min_{b,\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n (1 - y_n(\mathbf{w}^T\mathbf{z}_n + b))}_{\mathcal{L}(b,\mathbf{w},\alpha)} \right)$$

其中，上式括号里面的是对拉格朗日函数$L(b,w,\alpha)$计算最小值。那么根据梯度下降算法思想：最小值位置满足梯度为零。首先，令$L(b,w,\alpha)$对参数b的梯度为零：

$$\frac{\partial L(b,w,\alpha)}{\partial b} = 0 = -\sum_{n=1}^{N} \alpha_n y_n$$

也就是说，最优解一定满足$\sum_{n=1}^{N} \alpha_n y_n = 0$。那么，我们把这个条件代入计算max条件中（与$\alpha_n \geq 0$同为条件），并进行化简：

$$\max_{\text{all } \alpha_n \geq 0, \sum y_n \alpha_n = 0} \left(\min_{b,\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n (1 - y_n(\mathbf{w}^T\mathbf{z}_n)) \right)$$

这样，SVM表达式消去了b，问题化简了一些。然后，再根据最小值思想，令$L(b,w,\alpha)$对参数w的梯度为零：

$$\frac{\partial L(b,w,\alpha}{\partial w} = 0 = w - \sum_{n=1}^{N} \alpha_n y_n z_n$$

即得到：

$$w = \sum_{n=1}^{N} \alpha_n y_n z_n$$

也就是说，最优解一定满足$w = \sum_{n=1}^{N} \alpha_n y_n z_n$。那么，同样我们把这个条件代入并进行化简：

$$\max_{\text{all } \alpha_n \geq 0, \sum y_n \alpha_n = 0, \mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n} \left(\min_{b,\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n - \mathbf{w}^T\mathbf{w} \right)$$

$$\Longleftrightarrow \max_{\text{all } \alpha_n \geq 0, \sum y_n \alpha_n = 0, \mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n} -\frac{1}{2}\left\| \sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n \right\|^2 + \sum_{n=1}^{N} \alpha_n$$

这样，SVM表达式消去了w，问题更加简化了。这时候的条件有3个：

- all $\alpha_n \geq 0$
- $\sum_{n=1}^{N} \alpha_n y_n = 0$
- $w = \sum_{n=1}^{N} \alpha_n y_n z_n$

SVM简化为只有α_n的最佳化问题，即计算满足上述三个条件下，函数 $-\frac{1}{2}||\sum_{n=1}^{N} \alpha_n y_n z_n||^2 + \sum_{n=1}^{N} \alpha_n$最小值时对应的$\alpha_n$是多少。

总结一下，SVM最佳化形式转化为只与α_n有关：

$$\max_{\text{all } \alpha_n \geq 0, \sum y_n \alpha_n = 0, \mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n} -\frac{1}{2}\left\| \sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n \right\|^2 + \sum_{n=1}^{N} \alpha_n$$

其中，满足最佳化的条件称之为Karush-Kuhn-Tucker(KKT)：

if primal-dual optimal $(b, \mathbf{w}, \alpha)$,

- primal feasible: $y_n(\mathbf{w}^T\mathbf{z}_n + b) \geq 1$
- dual feasible: $\alpha_n \geq 0$
- dual-inner optimal: $\sum y_n \alpha_n = 0$; $\mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n$
- primal-inner optimal (at optimal all 'Lagrange terms' disappear):

$$\alpha_n(1 - y_n(\mathbf{w}^T\mathbf{z}_n + b)) = 0$$

—called **Karush-Kuhn-Tucker (KKT) conditions**, necessary for optimality [& sufficient here]

在下一部分中，我们将利用KKT条件来计算最优化问题中的α，进而得到b和w。

Solving Dual SVM

上面我们已经得到了dual SVM的简化版了，接下来，我们继续对它进行一些优化。首先，将max问题转化为min问题，再做一些条件整理和推导，得到：

standard hard-margin SVM dual

$$\min_{\alpha} \quad \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m \mathbf{z}_n^T\mathbf{z}_m - \sum_{n=1}^{N}\alpha_n$$

$$\text{subject to} \quad \sum_{n=1}^{N} y_n\alpha_n = 0;$$

$$\alpha_n \ge 0, \text{for } n = 1, 2, \ldots, N$$

(convex) QP of N variables & $N+1$ constraints, as promised

显然，这是一个convex的QP问题，且有N个变量α_n，限制条件有N+1个。则根据上一节课讲的QP解法，找到Q，p，A，c对应的值，用软件工具包进行求解即可。

optimal $\alpha = ?$

$$\min_{\alpha} \quad \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m \mathbf{z}_n^T\mathbf{z}_m - \sum_{n=1}^{N}\alpha_n$$

$$\text{subject to} \quad \sum_{n=1}^{N} y_n\alpha_n = 0;$$

$$\alpha_n \ge 0,$$

$$\text{for } n = 1, 2, \ldots, N$$

optimal $\alpha \leftarrow \text{QP}(\mathrm{Q}, \mathbf{p}, \mathrm{A}, \mathbf{c})$

$$\min_{\alpha} \quad \frac{1}{2}\alpha^T \mathrm{Q}\alpha + \mathbf{p}^T\alpha$$

$$\text{subject to} \quad \mathbf{a}_i^T\alpha \ge c_i,$$

$$\text{for } i = 1, 2, \ldots$$

- $q_{n,m} = y_n y_m \mathbf{z}_n^T\mathbf{z}_m$
- $\mathbf{p} = -\mathbf{1}_N$
- $\mathbf{a}_{\ge} = \mathbf{y}$, $\mathbf{a}_{\le} = -\mathbf{y}$;
  $\mathbf{a}_n^T = n$-th unit direction
- $c_{\ge} = 0$, $c_{\le} = 0$; $c_n = 0$

求解过程很清晰，但是值得注意的是，$q_{n,m} = y_n y_m z_n^T z_m$，大部分值是非零的，称为dense。当N很大的时候，例如N=30000，那么对应的Q_D的计算量将会很大，存储空间也很大。所以一般情况下，对dual SVM问题的矩阵Q_D，需要使用一些特殊的方法，这部分内容就不再赘述了。

- $q_{n,m} = y_n y_m \mathbf{z}_n^T \mathbf{z}_m$, often non-zero
- if $N = 30,000$, dense Q_D (N by N symmetric) takes > 3G RAM
- need **special solver** for
  - not storing whole Q_D
  - utilizing special constraints properly

  to scale up to large N

得到α_n之后，再根据之前的KKT条件，就可以计算出w和b了。首先利用条件$w = \sum \alpha_n y_n z_n$得到w，然后利用条件$\alpha_n(1 - y_n(w^T z_n + b)) = 0$，取任一$\alpha_n \neq 0$即$\alpha_n$>0的点，得到$1 - y_n(w^T z_n + b) = 0$，进而求得$b = y_n - w^T z_n$。

KKT conditions

if primal-dual optimal $(b, \mathbf{w}, \alpha)$,

- primal feasible: $y_n(\mathbf{w}^T \mathbf{z}_n + b) \geq 1$
- dual feasible: $\alpha_n \geq 0$
- dual-inner optimal: $\sum y_n \alpha_n = 0$; $\mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n$
- primal-inner optimal (at optimal all 'Lagrange terms' disappear):

$$\alpha_n(1 - y_n(\mathbf{w}^T \mathbf{z}_n + b)) = 0 \text{ (complementary slackness)}$$

- optimal $\alpha \Longrightarrow$ optimal $\mathbf{w}$? easy above!
- optimal $\alpha \Longrightarrow$ optimal b? a range from primal feasible & equality from **comp. slackness** if one $\alpha_n > 0 \Rightarrow b = y_n - \mathbf{w}^T \mathbf{z}_n$

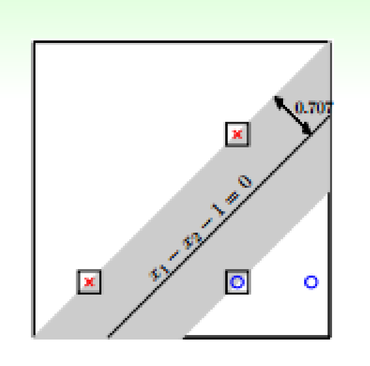
值得注意的是，计算b值，α_n>0时，有$y_n(w^T z_n + b) = 1$成立。$y_n(w^T z_n + b) = 1$正好表示的是该点在SVM分类线上，即fat boundary。也就是说，满足α_n>0的点一定落在fat boundary上，这些点就是Support Vector。这是一个非常有趣的特性。

Messages behind Dual SVM

回忆一下，上一节课中，我们把位于分类线边界上的点称为support vector（candidates）。本节课前面介绍了α_n>0的点一定落在分类线边界上，这些点称之为support vector（注意没有candidates）。也就是说分类线上的点不一定都是支

持向量，但是满足α_n>0的点，一定是支持向量。

- on boundary: 'locates' fattest hyperplane; others: **not needed**
- examples with $\alpha_n > 0$: on boundary
- call $\alpha_n > 0$ examples $(\mathbf{z}_n, y_n)$ **support vectors** ~~(candidates)~~
- SV (positive α_n) $\subseteq$ SV candidates (on boundary)



SV只由α_n>0的点决定，根据上一部分推导的w和b的计算公式，我们发现，w和b仅由SV即α_n>0的点决定，简化了计算量。这跟我们上一节课介绍的分类线只由“胖”边界上的点所决定是一个道理。也就是说，样本点可以分成两类：一类是support vectors，通过support vectors可以求得fattest hyperplane；另一类不是support vectors，对我们求得fattest hyperplane没有影响。

- only SV needed to compute **w**: $\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n = \sum_{\text{SV}} \alpha_n y_n \mathbf{z}_n$
- only SV needed to compute b: $b = y_n - \mathbf{w}^T \mathbf{z}_n$ with any SV $(\mathbf{z}_n, y_n)$

回过头来，我们来比较一下SVM和PLA的w公式：

SVM	PLA
$\mathbf{w}_{\text{SVM}} = \sum_{n=1}^{N} \alpha_n (y_n \mathbf{z}_n)$	$\mathbf{w}_{\text{PLA}} = \sum_{n=1}^{N} \beta_n (y_n \mathbf{z}_n)$
α_n from **dual solution**	β_n by **# mistake corrections**

我们发现，二者在形式上是相似的。w_{SVM}由fattest hyperplane边界上所有的SV决定，w_{PLA}由所有当前分类错误的点决定。w_{SVM}和w_{PLA}都是原始数据点$y_n z_n$的线性组合形式，是原始数据的代表。

$\mathbf{w}$ = linear combination of $y_n\mathbf{z}_n$

- also true for GD/SGD-based LogReg/LinReg when $\mathbf{w}_0 = \mathbf{0}$
- call $\mathbf{w}$ 'represented' by data

总结一下，本节课和上节课主要介绍了两种形式的SVM，一种是Primal Hard-Margin SVM，另一种是Dual Hard_Margin SVM。Primal Hard-Margin SVM有$\hat{d}+1$个参数，有N个限制条件。当$\hat{d}+1$很大时，求解困难。而Dual Hard_Margin SVM有N个参数，有N+1个限制条件。当数据量N很大时，也同样会增大计算难度。两种形式都能得到w和b，求得fattest hyperplane。通常情况下，如果N不是很大，一般使用Dual SVM来解决问题。

Primal Hard-Margin SVM	Dual Hard-Margin SVM
$\min_{b,\mathbf{w}} \ \frac{1}{2}\mathbf{w}^T\mathbf{w}$ sub. to $y_n(\mathbf{w}^T\mathbf{z}_n + b) \geq 1$, for $n = 1, 2, \ldots, N$	$\min_{\boldsymbol{\alpha}} \ \frac{1}{2}\boldsymbol{\alpha}^T \mathrm{Q}_\mathrm{D} \boldsymbol{\alpha} - \mathbf{1}^T\boldsymbol{\alpha}$ s.t. $\mathbf{y}^T\boldsymbol{\alpha} = 0$; $\alpha_n \geq 0$ for $n = 1, \ldots, N$
• $\tilde{d}+1$ variables, N constraints —suitable when $\tilde{d}+1$ small	• N variables, $N+1$ simple constraints —suitable when N small
• physical meaning: locate specially-scaled $(b, \mathbf{w})$	• physical meaning: locate SVs $(\mathbf{z}_n, y_n)$ & their α_n

both eventually result in optimal $(b, \mathbf{w})$ for fattest hyperplane

$$g_{\text{SVM}}(\mathbf{x}) = \text{sign}(\mathbf{w}^T\boldsymbol{\Phi}(\mathbf{x}) + b)$$

这节课提出的Dual SVM的目的是为了避免计算过程中对$\hat{d}$的依赖，而只与N有关。但是，Dual SVM是否真的消除了对$\hat{d}$的依赖呢？其实并没有。因为在计算$q_{n,m} = y_n y_m z_n^T z_m$的过程中，由z向量引入了$\hat{d}$，实际上复杂度已经隐藏在计算过程中了。所以，我们的目标并没有实现。下一节课我们将继续研究探讨如何消除对$\hat{d}$的依赖。

$$\min_{\alpha} \quad \frac{1}{2}\alpha^T Q_D \alpha - \mathbf{1}^T \alpha$$

$$\text{subject to} \quad \mathbf{y}^T \alpha = 0;$$

$$\alpha_n \geq 0, \text{for } n = 1, 2, \ldots, N$$

- N variables, $N + 1$ constraints: **no dependence on $\tilde{d}$?**
- $q_{n,m} = y_n y_m \mathbf{z}_n^T \mathbf{z}_m$: inner product in $\mathbb{R}^{\tilde{d}}$ —$O(\tilde{d})$ via naïve computation!

总结

本节课主要介绍了SVM的另一种形式：Dual SVM。我们这样做的出发点是为了移除计算过程对$\hat{d}$的依赖。Dual SVM的推导过程是通过引入拉格朗日因子α，将SVM转化为新的非条件形式。然后，利用QP，得到最佳解的拉格朗日因子α。再通过KKT条件，计算得到对应的w和b。最终求得fattest hyperplane。下一节课，我们将解决Dual SVM计算过程中对$\hat{d}$的依赖问题。

注明：

文章中所有的图片均来自台湾大学林轩田《机器学习技法》课程